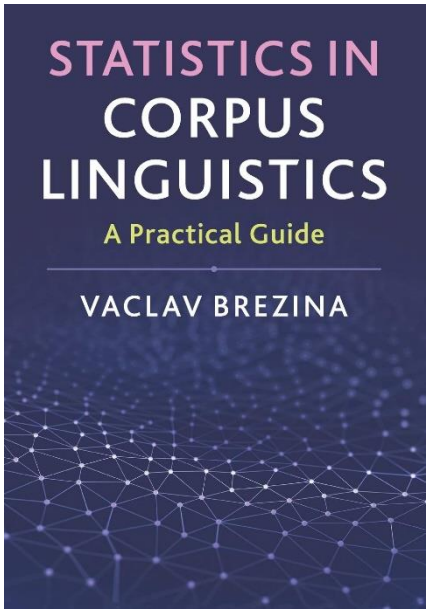
Statistics in Corpus Linguistics: A Practical Guide

Vaclav Brezina

Cambridge, Cambridge University Press, 2018 pp. 296



Review by Valeria Franceschi*

Corpus linguistics is a steadily growing field that allows scholars to investigate language by means of massive electronically stored databases of texts that are interrogated through software. Research carried out on corpora ranges from terminology to sociolinguistics to discourse analysis. Empirical results from the analysis are employed to support—or disprove—hypotheses about language. Such output - frequencies of words and phrases in corpora and their distribution patterns—is considered quantitative data, and statistical procedures are necessary to work with and interpret such quantitative data (3; McEnery and Hardie 2011). However, many linguists are not trained in such statistical techniques or even familiar with the statistical thinking behind the methods used in corpus research, and how these may vary according to different research designs and research questions. With this volume, Brezina sets out to "help readers understand key principles of statistical thinking in order to be able to make informed decisions about the applications of particular statistical techniques" (xvii). And indeed, this volume fills an existing gap by providing an accessible approach to theoretical and practical basics to corpus scholars who want to familiarize themselves with relevant statistical procedures.

The volume consists of 8 chapters, each of which deals with a different topic related to statistics in corpus linguistics. Each chapter is made up of multiple parts, effectively guiding the reader through the theoretical concepts covered and their application to corpus analysis while fostering reasoning and understanding through reflection activities. Chapters open with a brief "What is this chapter about?" section which anticipates the contents of the chapter and sets forward a number of questions that the different sections of

* *Valeria Franceschi is currently a junior researcher in English Language and Translation at the Department of Foreign Languages and Literatures of the University of Verona, where she teaches at undergraduate and graduate level. She has recently published* Exploring Plurilingualism in Fan Fiction: ELF Users as Creative Writers *(Cambridge Scholars Publishing, 2017). Lately, her research in ELF has focused on plurilingual practices and on the business context (BELF). In addition to ELF, her research areas of interest include digital communication and corpus linguistics.*

the chapter will answer. An expository part follows, where definitions and methods are illustrated in detail as well as how results should be reported following good practice. The expository part closes with a case study where a practical application of the methods introduced in each chapter is shown. Readers can then practice what they have just learned by going through the comprehensive exercise section. Chapters close with a "Things to Remember" box that recaps the major points discussed and additional bibliographical references for more advanced reading on the topics. In addition to the material in the paper book, the volume also has a companion website[1]—Lancaster Stats Tools online—where readers can find a range of other useful materials, including exercise keys, and the tools to perform the necessary calculations and create the visual representations required in the activities.

Looking at chapter content more in detail, Chapter 1 acts as an introductory chapter, as it draws a connection between the "basic principles of statistical thinking" (1) and corpus linguistics. A section of statistical terminology and fundamental concepts, such as inferential statistics and statistical significance, is followed by good practice indications in research and corpus design, and by a section on different ways of visualizing data (e.g. bar charts, boxplots and scatterplots). The expository part of the chapter closes with a case study on adjective use in academic discourse versus fiction.

Chapter 2, which focuses on vocabulary, starts by providing definitions to terms commonly used in corpus linguistics when dealing with words (token, type, lemma, lexeme), and how statistics on these numbers are reported. The importance of dispersion is then highlighted, with the illustration of different dispersion measures and how such measures are reported. Lexical diversity is then introduced, and once again the chapter closes with the stage-by-stage illustration of a case study, this time on the prominence of *weather*-related content words in British English.

In Chapter 3 the book explores semantics and discourse, with a focus on collocation and keywords. After covering the basic terminology related to collocation, the author introduces multiple association measures and illustrates the results each of them yields. Different association measures are used according to the type of collocation the scholar wants to investigate. Collocation networks are also explained. Issues related to the investigation of keywords are introduced in the second part of the chapter, from choosing a reference corpus to corpus comparison and inter-rater agreement measures. The chapter closes with a case study using collocation and keyword for the purpose of discourse analysis.

Chapter 4 covers lexico-grammar analysis, introducing two different approaches to this type of investigation: whole corpus design, which includes the whole corpus or a large subcorpus as a unit of analysis, and linguistic feature design, which "focuses on the linguistic feature as a single observation (case)" (22). This second approach is then explored more in depth, and a range of statistic techniques for data analysis is illustrated, with a long section focusing on the multiple steps of logistic regression. An application of such technique is then demonstrated in the case study section of the chapter.

Register variation is explored in Chapter 5. The first statistical concept covered here is that of correlation and the different ways in which it can be measured according to the variables under investigation. Following discussion on correlation, hierarchical agglomerative cluster analysis is introduced as a way to classify "'objects' (words, sentences, texts or speakers etc.) that can be characterized using multiple linguistic features" (151) when working with non-diachronic data. The chapter continues with a discussion of Multidimensional Analysis (MD), a procedure used when working with many linguistic variables to determine how they co-occur, in order to identify the underlying principles of functional variation in the data. A case study on registers in New Zealand closes the chapter before the exercise section.

Chapter 6 deals with sociolinguistics and stylistics, both in individual and social variation. It provides readers with a range of statistical procedures that can be used to investigate how individual speakers and groups of speakers can be compared on the basis of specific social factors. Group comparison can be carried out with multiple tests: t-test, ANOVA, Mann-Whitney U Test, Kruskal-Wallis test, whereas correspondence analysis is employed to identify the style of individual speakers. Mixed-effect models are instead discussed in relation to sociolinguistic variation in specific linguistic contexts. The case study for this chapter demonstrates how the author found the solution to a sociolinguistic 'riddle' and identified the speaker of the sample he had been sent.

[1] http://corpora.lancs.ac.uk/stats. Last Visited August 23, 2019.

Chapter 7 focuses on diachronic data and on the statistical procedures used in longitudinal studies. The chapter opens with basic definitions of diachronic corpora and the aspects to consider when working with such data. Differences across time have to be tested for significance with a non-parametric test called bootstrap. Neighboring cluster analysis, either hierarchical or variability-based, is instead used to "[categorize] individual temporal data points […] into larger historical periods based on the similarity in language use" (237). Following this, the peaks and throughs method is explained alongside an extension on this technique, Usage Fluctuation Analysis (UFA). In this chapter, the procedures are applied to variation in the use of color words during the 17th century.

Chapter 8 acts as a conclusive chapter. First, it underlines the ten key principles of statistical thinking; second, it introduces meta-analysis, a statistical technique employed to bring together and systematically combine results from different studies and their systematic combination. In addition to illustrating the workings of meta-analysis, this chapter argues for the importance of reporting standardized effect size measures and provides a range of frequently-used effect size measures and illustrates how these measures are interpreted.

The structure of the chapters, the range of topics discussed, its accessibility and focus on the practical application of the illustrated principles make it an essential tool for any scholar working or planning to work with quantitative data in corpus linguistics. The author guides the reader successfully through the material by interspersing exposition with reflection activities that foster thinking and comprehension of concepts that may not always be straightforward for scholars and students with a humanities background and no previous training in mathematics or statistics. The extensive exercise sections make for plenty of practice with the statistical procedures; the calculation tools provided on the companion website ensure that readers have the necessary instruments to successfully perform all the steps required by the suggested exercise. The same website also provides additional material for teachers and students—indeed, in addition to being a valuable book for self-study, it may also be used as a textbook for more advanced students in corpus linguistics.

Works cited

McEnery, Tom and Andrew Hardie. *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press, 2011.